UNIVERSITÉ LAVAL

Faculté des sciences de l’agriculture et de l’alimentation
Département de phytologie

Understanding and predicting pesticide use on golf courses using deep machine learning

Interim report

March 2022

Guillaume Grégoire, Ph.D.
Assistant professor, Département de phytologie

Executive summary

Golf course maintenance requires the use of several inputs, such as pesticides and fertilizers, that can be harmful to human health or the environment. Understanding the factors associated with pesticide use on golf courses could help them reduce their reliance on these products. In this paper, a database of about 14000 pesticide applications in the province of Québec, was used to develop a novel hybrid machine learning approach to predict pesticide use on golf courses. This proposed model, called RF-SVM-GOA, was created by coupling Support Vector Machine (SVM) with Random Forest (RF) and Grasshopper Optimization Algorithm (GOA). Five different dependent variables including region, golf course ID, number of holes, year, and treated area were considered as input variables. The experimental results confirmed that the developed hybrid RF-SVM-GOA technique was able to estimate the active ingredient total (AIT) with a high level of accuracy (R = 0.99; MAE = 0.84; RMSE = 0.84; NRMSE = 0.04). A sensitivity analysis was performed to find the most effective variables in AIT forecasting. The results indicated that the treated area is the most effective variable in AIT forecasting. The results of the current study could be one step forward to sustainable golf course management. Next steps will be to integrate weather data (air temperature and precipitations) to the model and predict pesticide use under different climate change scenarios.

Introduction

Golf courses are open green spaces usually located in urban or peri-urban settings and provide several benefits to the community. However, inputs such as pesticides and fertilizers used to maintain their playing surfaces can also have detrimental effects on the environment and human health. To address those concerns about pesticide use and their impact on human health and the environment, the province of Québec requires, since 2003, all golf courses to submit a triennial pesticide reduction plan, signed by a certified agronomist, to its Ministry of Environment (MOE). This plan must also include pesticide reduction objectives for the following three years, and methods that will be implemented to achieve those objectives. Since several unpredictable factors can affect pest pressure and pesticide use, it can be difficult for golf managers and agronomists to define objectives that will be realistic, but that will also result in the greatest possible reductions.

Since the adoption of the Pesticide Management Code in 2003, the Québec Ministry of Environment (MOE) has compiled a database of all pesticides applications on golf courses in the province. While the MOE publishes triennial reports on pesticide use, this database has never been subjected to a deep analysis to understand the impact of this regulation as well as the different factors influencing pesticide use. Recent advances in machine learning (ML) have made it possible to develop algorithms capable of identifying hidden trends in large datasets, and this process has been previously used in agriculture in the past. Some of the well-known machine learning techniques are artificial neural networks, support vector machines and decision trees.

The main objective of the current project was to use ML techniques on the pesticide application database to understand and predict pesticide use on golf courses in Québec. The project was divided in three parts:

1- Developing a model to analyze the pesticide database and identify factors associated with pesticide use. Status: completed

2- Add weather data to the model and predict pesticide use according to different climate change scenarios. Status: in progress

3- Select a subset of golf courses with contrasting pesticide use and investigate specific factors and practices influencing pesticide use. Status: in progress

For Part 1 specifically, the objectives were to develop a ML algorithm that could: 1- analyze a database a subset of about 14 000 pesticide applications made on golf courses in the province of Québec, Canada; 2- identify golf course characteristics associated with pesticide use; and 3- to predict future pesticide use to better guide golf courses in their pesticide reduction objectives.

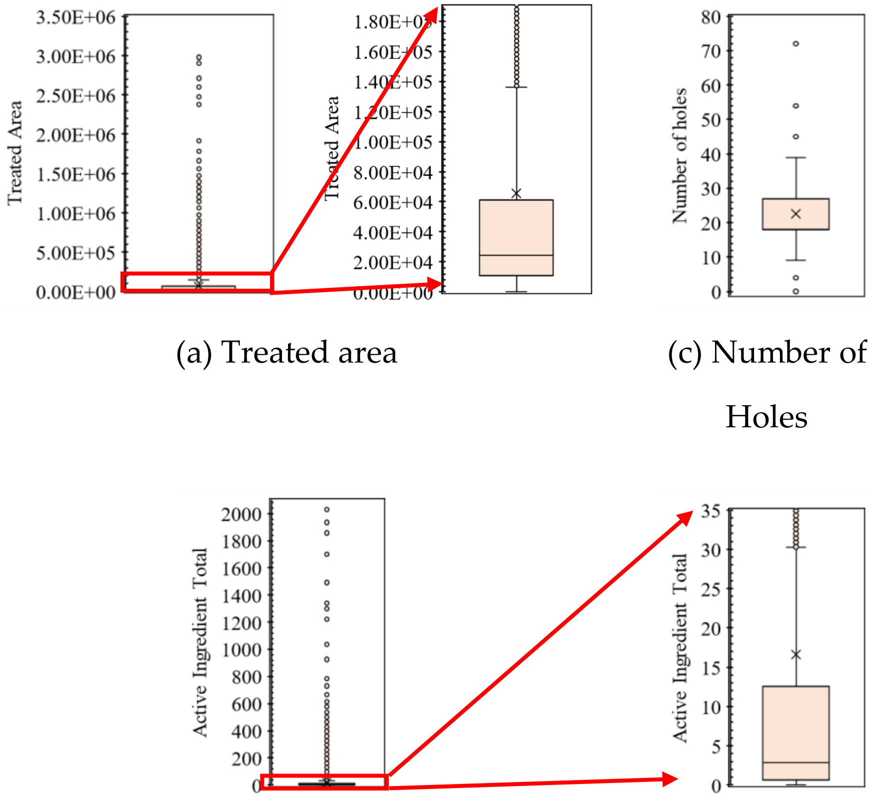
Methodology

The database provided by the MOE contains 53 054 pesticide applications made on golf courses in Québec from 2003 to 2017. The database was first treated to remove missing values and duplicates, and filter aberrant data. Pesticide applications where the quantity of active ingredient applied was more than 3 times the maximum label rate were considered as outliers and thus removed from the database. This resulted in a final database containing 41 456 pesticide applications.

To develop the ML model, we used a subset of 13 220 unique pesticide applications randomly selected. This subset was used for the development of the ML model, using five input variables (treated area (TA), administrative region (AR), golf course ID (GFI), number of holes (NH) and year(Y)) and one output variable (AIT). Each golf course with its fixed number of holes was attributed a given ID number in the MOE database (golf course ID). The treated area (in m^2) corresponds to the area treated with a given pesticide, while the region corresponds to the number of the administrative region (number 1 to 17) of the province of Quebec where the golf course is located. For each application, active ingredient total (AIT) was calculated as $AIT = Q \times C$, where Q is the quantity of pesticide applied in kg or L and C is the concentration of the active ingredient in the applied pesticide in percentage. For modelling purposes, 70% of the 13220 samples were randomly selected as calibration samples and the rest were considered as validation samples.

A large variation (i.e. difference between minimum and maximum value) was observed for TA and AIT values (Figure 1). Given that there are large numbers of samples in the high ranges of these variables, they cannot be considered as outliers. Due to this large variation in AIT as the outcome, the machine learning model faces a significant challenge for adjusting the model hyper parameters and its training to achieve a model with a good prediction performance for all range of samples. Thus, we needed to develop a hybrid model by coupling Support Vector Machine (SVM) with Random Forest (RF) and Grasshopper Optimization Algorithm (GOA).

Random Forest (RF) is a powerful ensemble classifier that is robust against overfitting. RF employs a set of Classification and Regression Trees (CART) methodologies to map the independent input variables to the dependent output variable(s). The trees in RF are generated using a bagging approach by the selection of a subset of training samples with replacement. Indeed, some samples may be chosen multiple times, while some may not be chosen at all.



(a) Treated area

(c) Number of

Holes

(b) Active Ingredient Total

Figure 1. The box plot of the independent input variables and dependent output variable (a: Treated Area (TA); b: Active Ingredient Total (AIT); c: Number of Holes (NH)

Support vector machine (SVM) is a well-known robust, reliable, and efficient supervised machine learning technique for classification and regression problems. The modelling framework in this method is defined based on statistical learning. SVM finds a special type of linear model that maximizes hyperplane margin. Maximizing hyperplane margin leads to maximizing the separation between classes. Indeed, the training points closest to the maximum hyperplane margin are called support vectors, which are used to define the margin between classes. If the data can be separated linearly, linear machines are used to generate an optimal level that separates the data without error and with the maximum distance between the plane and the nearest training points (support vectors).

Grasshopper Optimization Algorithm (GOA) is an evolutionary nature-inspired algorithm that imitates the social interactions and demeanor of the grasshopper searching for food in nature to define a mathematical model in solving optimization problems. The grasshopper's life cycle includes egg, nymph, and adult stages. The swarming behavior of grasshoppers occurs at both nymph and adult stages. The small steps and low movement of grasshoppers occur in the larval phase while the abrupt movement and long-range movement takes place in adulthood. In addition to these two features of the grasshoppers' swarming, the search for a food source is another vital feature. Consequently, moving abruptly and locally as exploration and exploitation search processes as well as target seeking, are naturally performed by grasshoppers. Therefore, the natural behavior of grasshoppers can be defined mathematically, and the resulting equation can be applied to large datasets.

The developed hybrid method is a new coupling of SVM with two different methods: RF as a classifier to overcome the high range of samples and GOA as an optimization algorithm to find the optimal value of the SVM. The modelling process of the developed RF-SVM-GOA model is done through an integrated computer program in the MATLAB environment. Figure 2 indicates the flowchart of the developed hybrid techniques in AIT forecasting. The first step is started by categorizing all samples into calibration and validation data. Using calibration samples, different random vectors are generated through the training phase based on the defined framework for RF and employed to build multiple decision trees (DT). The created DTs are combined to find the final trees. For each selected tree, random values for a pre-defined swarm are initialized and the fitness function is calculated for this swarm. After that, the support vector machine is implemented, and the cost function is recalculated to find the best search agent. This process is repeated to reach all search agents at different iteration numbers as well as for all trees to find the final model for AIT forecasting.

According to the stochastic nature of AIT, considering only one criterion to assess the performance of each model is not sufficient. Therefore, a group of statistical indices including correlation coefficient (R), mean absolute error (MAE), root mean square error (RMSE), and normalized RMSE (NRMSE), is applied to check the performance of the developed models in AIT forecasting. Considering the mentioned indices simultaneously is sufficient to evaluate the efficiency of a model.

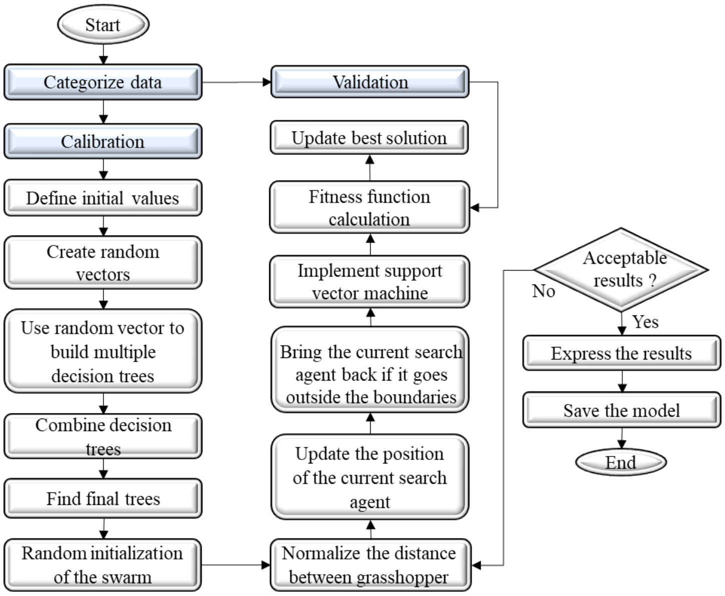

Start
Categorize data
Calibration
Define initial values
Create random vectors
Use random vector to build multiple decision trees
Combine decision trees
Find final trees
Random initialization of the swarm
Normalize the distance between grasshopper
Update the position of the current search agent
Bring the current search agent back if it goes outside the boundaries
Implement support vector machine
Fitness function calculation
Update best solution
Validation
Acceptable results ?
No
Yes
Express the results
Save the model
End

Figure 2. The flowchart of the developed hybrid techniques in AIT forecasting

Results

Figure 3 demonstrates the scatter plot of the hybrid model developed for AIT forecasting. The model performed well, with statistical indices of R = 0.99; MAE =0.84; RMSE = 0.84; NRMSE = 0.04.

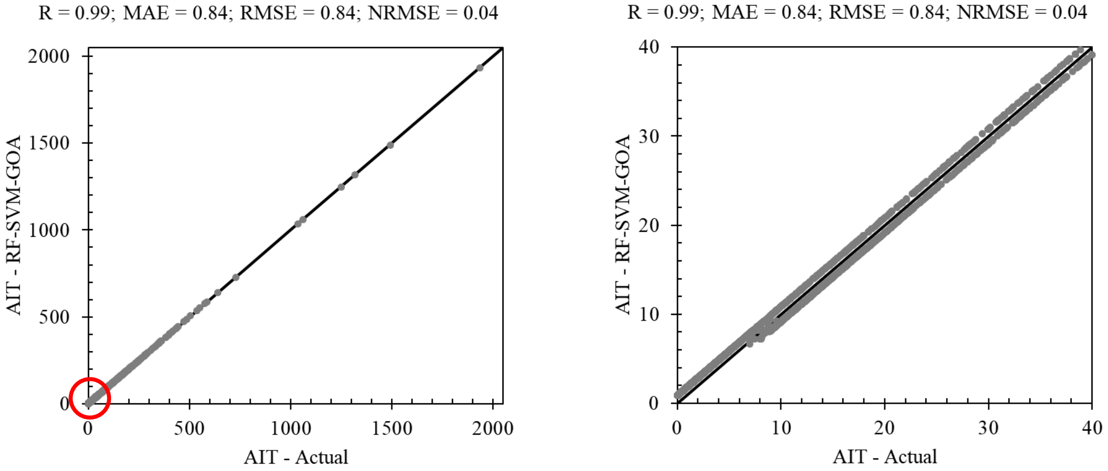


Figure 3. Scatter plot of the resulting hybrid RF-SVM-GOA model for AIT forecasting. Plot on the right represent a scaled-up view of the region circled in red on the left plot.

The effect of each independent input variable on AIT forecasting using the developed hybrid technique is examined in Table 1. According to this table, removing Treated Area (TA) from the model inputs (Model 2) significantly reduces the modelling accuracy compared to Model 1. Furthermore, the correlation coefficient of Model 2 is less than 1% of the value of this index in Model 1, while the RMSE, NRMSE, and MAE indices in Model 2 are more than 362, 570, and 570 (respectively) times the value of each index in Model 1. Removing Y and GCI as inputs variables remarkably also decrease the AIT modelling accuracy. Although model 5 and 6 have poor performance in AIT estimation, they have higher accuracy than model 2. Similar to the three variables GCI, TA, and Y, not using the two variables NH and R (models 5 and 6) also affect the modelling results. However, the importance of these two variables (i.e., NH and AR) are lower than the others. The correlation coefficients of models 5 and 6 are 47% lower than Model 1. The value of the RMSE, NRMSE and MAE for models 5 and 6 are more than 3, 59, and 59 times of their values at Model 1. TA, Y, GCI, NH, and AR are ranked first to fifth (respectively) for ranking the effect of different input variables.

Table 1. The effect of each independent input variable on the AIT forecasting using the developed hybrid technique.

AR	NH	GCI	Y	TA	Model No.	R^2	RMSE	NRMSE	MAE
•	•	•	•	•	Model 1	0.999	0.839	0.843	0.044
•	•	•	•		Model 2	0.003	303.782	480.594	25.072
•	•	•		•	Model 3	0.011	73.952	261.179	13.625
•	•		•	•	Model 4	0.020	54.141	204.369	10.661
•		•	•	•	Model 5	0.522	2.563	49.765	2.596
	•	•	•	•	Model 6	0.522	2.563	49.766	2.596

AR = Administrative Region, NH = Number of holes, GCI = Golf course ID, Y = Year, TA = Treated area

This model was developed using data that was available under the current regulations in the province of Québec. However, other factors affecting pesticide use on golf courses are not considered in these regulations. For example, other research has showed a correlation between pesticide use and golf course economic data such as revenue per hectare and maintenance budget. Other factors, such as superintendent experience and level of education, golfers' expectations and local environmental characteristics could also affect pesticide use but are not included in the applied model in the current study.

This model could be helpful to help Québec golf managers to better predict their pesticide use and to be more efficient in fixing their pesticide reduction objectives as required by the current regulations. Agronomists and regulatory authorities could also use it to better refine the process by which golf courses fix and achieve their objectives in pesticide reduction. For example, favourable economic measures could be implemented for golf courses that use fewer pesticides than what was predicted by the model.

Knowledge transfer activities

Results from this project were presented (virtually) at the annual meeting of the Québec society for plant protection in September 2021 and to the Québec Golf Superintendents Association annual meeting in February 2022. A scientific paper was submitted to the journal Agriculture (MDPI) in February 2022 and is now in the revision process.

Next steps

We are currently in the process of integrating weather data (air temperature and precipitations) to the model. We will consider the both the impact of summer weather (i.e. May to October) and winter weather (November to April) in our analysis. We will then be able to simulate different climate change scenarios and predict pesticide applications under each one.

We also started pooling golf courses based on their pesticide use (low, medium and high). One this is completed, we will identify and contact specific golf courses in each category to investigate the specific factors and practices they put in place to achieve their objectives in pesticide reduction. A summer student will be hired in the next few weeks to specifically work on this part of the project.